Pre-class activities
Support Vector Machines and Kernels

Emmanuel Rachelson

1 Quadratic programming

Suppose a collection of N pairs $\{(x_i, y_i)\}_{i=1..N}$ with $x_i \in \mathbb{R}^n$ and $y_i \in \{-1; 1\}$ and consider the following optimization problem:

$$\min_{w \in \mathbb{R}^n} \frac{1}{2}\|w\|^2$$
$$s.t.\ \forall j = 1...N, \quad y_i \left(w^T x_i\right) \leq 1$$

☞ Recall the name of this type of optimization problem.

It is a continous, twice differentiable optimization problem over the w variables. Moreover, the objective function is quadratic while the constraints are linear: it is a Quadratic Programming (QP) problem.

☞ Write down the problem's Lagrangian.

Let α_i be the Karush-Kuhn-Tucker coefficient associated to the constraint $y_i \left(w^T x_i\right) - 1 \leq 0$. Then:

$$L(w, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i \left(y_i \left(w^T x_i\right) - 1\right)$$

☞ Why are the constraints qualified?

The constraints are linear, thus qualified.

☞ Write Karush-Kuhn-Tucker's first order conditions. Recall what they mean.

KKT first order conditions provide a necessary optimality condition. They state that:

$$w^* \text{ is a solution} \Rightarrow \exists \alpha^* \in \mathbb{R}^{+N} \ s.t. \begin{cases} \frac{\partial L}{\partial w}(w^*, \alpha^*) = 0 \\ \forall i \in [1, N], \quad \alpha_i \left(y_i \left(w^T x_i\right) - 1\right) = 0 \end{cases}$$

In other words, if one can find a pair $(w^*, \alpha^*) \in \mathbb{R}^n \times \mathbb{R}^{+N}$ that verifies the right-hand side equations of the line above, then it is a candidate optimal solution.

In this particular case, one has a convex objective function and convex constraints, so KKT first order necessary condition is also a sufficient condition.

The condition writes:

$$\left\{\begin{array}{ll} w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 & \\ \forall i \in [1, N], & \alpha_i \left(y_i \left(w^T x_i \right) - 1 \right) = 0 \\ \forall i \in [1, N], & \alpha_i \geq 0 \end{array}\right.$$

☞ Recall the duality theorem in Differentiable Optimization and write the dual form of the above optimization problem.

The duality theorem states that if the objective function and constraints are convex, and if the constraints are qualified, then the solution to the dual problem $\sup_{\alpha \geq 0} \inf_w L(w, \alpha)$ *is the solution to the optimization problem.*

By using the first equation in the first order conditions above, one obtains $w = \sum_{i=1}^{N} \alpha_i y_i x_i$. *Consequently, the dual function* $L_D(\alpha)$ *writes:*

$$L_D(\alpha) = \frac{1}{2} \left\| \sum_{i=1}^{N} \alpha_i y_i x_i \right\|^2 - \sum_{i=1}^{N} \alpha_i \left(y_i \left(\left(\sum_{j=1}^{N} \alpha_j y_j x_j \right)^T x_i \right) - 1 \right)$$

The first term becomes:

$$\begin{aligned} \frac{1}{2} \left\| \sum_{i=1}^{N} \alpha_i y_i x_i \right\|^2 &= \frac{1}{2} \left(\sum_{i=1}^{N} \alpha_i y_i x_i \right)^T \left(\sum_{j=1}^{N} \alpha_j y_j x_j \right) \\ &= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \end{aligned}$$

While the second term yields:

$$\sum_{i=1}^{N} \alpha_i \left(y_i \left(\left(\sum_{j=1}^{N} \alpha_j y_j x_j \right)^T x_i \right) - 1 \right) = - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \alpha_i$$

And thus:

$$L_D(\alpha) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \alpha_i$$

Finally, solving the dual problem boils down to solving:

$$\sup_{\alpha \geq 0} \left[-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \alpha_i \right]$$

2 Insensitive Least squares regression

Suppose a collection of N points $\{(x_i, y_i)\}_{i=1..N}$ with $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ and consider the associated regression problem. Suppose also that a family of functions $\Phi = \{\phi_j\}_{j=1..p}$, $\phi_j : \mathbb{R}^n \to \mathbb{R}$ is provided and that the solution to the regression problem should lie in the space spanned by Φ (that is, the regression function f should have the form $f = \sum_{j=1}^{p} w_j \phi_j$).

☞ Write the regression problem as a least squares minimization problem.

Under the least squares fitting criterion, the regression problem can be written:

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} w_j \phi_j(x_i) \right)^2$$

Or in vector form:

$$\min_{w \in \mathbb{R}^p} \left\| y - w^T \phi(x) \right\|^2$$

Consider the data plotted in Figure 1.

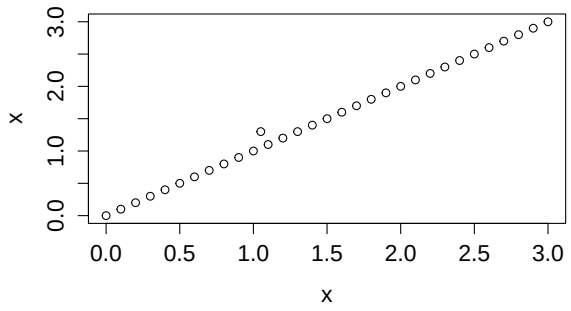

Figure 1: Regression data

Apart from the outlier at $x = 1.05$, the data fits the $y = x$ relation perfectly. In this case, advanced feature functions do not seem necessary, so the jth feature is actually the jth component of x. Thus the problem is a simple $y = w^T x$ regression problem.

☞ Can you think of a formulation of the regression problem that would be robust to noise? For example, that would discard any noisy point that stays inside a "tube" of width ϵ around the inferred function?

3 The trick of the additional dimension

Consider the following test data where x is a voltage measurement and y indicates whether an electronic component failed under that voltage:

x	0.3	0.7	1.1	1.8	2.5	3.0	3.3	3.5	3.7
y	-1	-1	-1	1	1	1	1	-1	-1

☞ Figure 3 shows a graphical display of the above data set. One wishes to linearly separate the data. What is the (very naive) general form of a linear classifier on this data ? What is the best training error one can obtain with such a classifier ?

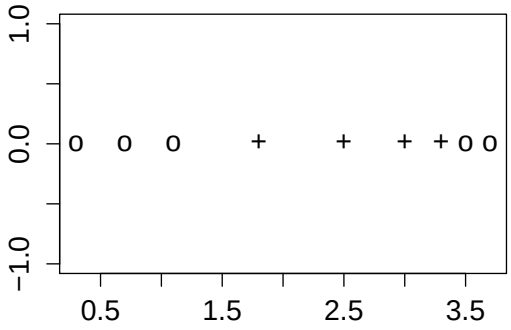

Figure 2: Raw measurements

A linear classifier splits the data based on the simple rule $class = sign(w^T x)$. *In this case, since* x *is unidimensional,* w *is a scalar.*

Based on Figure 3, the best training error one can obtain is 2 misclassified examples out of 9 (22%).

A smart engineer decides to plot the same data but enriches the description by adding a second axis representing $(2 - x)^2$. The data set becomes:

$z_1 = x$	0.3	0.7	1.1	1.8	2.5	3.0	3.3	3.5	3.7
$z_2 = (2 - x)^2$	2.89	1.69	0.81	0.04	0.25	1	1.69	2.25	2.89
y	-1	-1	-1	1	1	1	1	-1	-1

The new graphical representation is displayed on Figure 3.

☞ Did that operation seem to help the linear classification task? What lessons can one draw?

Figure 3 presents data that is linearly separable. Thus, the introduction of $z = (z_1, z_2)$ *allowed to transform a data set which was not linearly separable into*

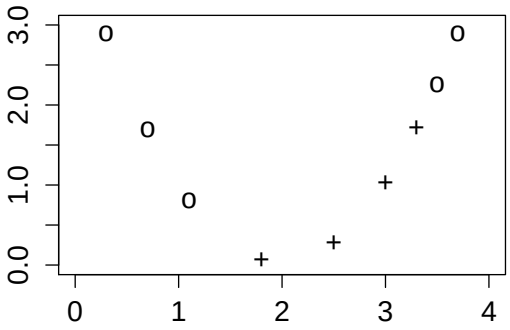

Figure 3: Enriched representation

a dataset that was. This illustrates the importance of feature engineering for any machine learning task. Furthermore, it illustrates the interest of enriching the data representation in order to "send" data such as x in a space of larger dimension (such as the space of z) where its representation is easier.